CONTENTS

More detailed tables of contents are to be found within the various parts of the compendium. The following provides merely an overview.

ACKNOWLEDGMENTS

The HIV Sequence Database and Analysis Project is funded by the Vaccine and Prevention Research Program of the AIDS Division of the National Institute of Allergy and Infectious Diseases (Dr. James Bradac, Project Officer) through an interagency agreement with the U.S. Department of Energy.

We thank the many researchers who have made their sequences available prior to publication.

The photograph on the cover of this compendium of the late Howard Temin was taken by Gregory Anderson and was kindly provided by Bette Sheehan of the University of Wisconsin.

INTRODUCTION

This compendium and the accompanying floppy diskettes are the result of an effort to compile and rapidly publish all relevant molecular data concerning the human immunodeficiency viruses (HIV) and related retroviruses. The scope of the compendium and database is best summarized by the five parts that it comprises: (I) Nucleic Acid Alignments and Sequences; (II) Amino Acid Alignments; (III) Analysis; (IV) Related Sequences; and (V) Database Communications. Information within all the parts is updated at least twice in each year, which accounts for the modes of binding and pagination in the compendium.

While this publication could take the form of a review or sequence monograph, it is not so conceived. Instead, the literature from which the database is derived has simply been summarized and some elementary computational analyses have been performed upon the data. Interpretation and commentary have been avoided insofar as possible so that the reader can form his or her own judgments concerning the complex information. In addition to the general descriptions below of the parts of the compendium, the user should read the individual introductions for each part.

Part I. Nucleic Acid Alignments and Sequences. Annotated nucleic acid sequences of certain HIVs and SIVs are presented in a form close to that of the GenBank Sequence Library. Our few modifications of standard GenBank format were instituted to better serve the particular community for which this database is intended. Beginning in 1995, most sequences are not presented but rather are catalogued, in order to conserve space; the full formatted GenBank entries of these sequences are located on the Web site (*http://hiv-web.lanl.gov*) and the database FTP server.

The LOCUS name or identifier of an entry may differ slightly from that found in the GenBank or EMBL libraries, but the ACCESSION numbers are identical for entries in all (four) nucelotide sequence databases. Thus each entry is universally and uniquely traceable. Sequences may also be described by COMMON NAMES.The SOURCE line provides information, when available, about the infectivity or biological activity of the molecular clone from which a sequence has been derived. REFERENCES are limited to literature or personal communications having authority for the original sequence data; references that review sequence information, or that shed light upon the function or variation of coding and regulatory sequences, are listed in Part V.

Entries in Part I are annotated within the sequence, while their GenBank or EMBL-formatted versions on the floppy diskettes make use of FEATURES tables. The hard-copy annotation includes coding regions, regulatory structures, splice sites, and other features of functional significance. The authority for this annotation is largely invariance, the recurrence of patterns such as TATAA and AATAAA. Although our practice has been to conservatively annotate, we caution the user against docility: sequence information regarding transcripts, for example, is far from certain or complete at this time. Part I is divided into three subsections, A, B, and C, concerned with HIV-1s, HIV-2/SIVs, and SIVAGMs, SIVMND and SIVSYK.

Part II. Amino Acid Alignments. This section contains in alignment the amino acid sequences (mostly full-length) of all known coding regions and open reading frames of HIV-1, HIV-2, and SIVs. Beginning in 1993, Part II is broken down into subsections A, B and C, for HIV-1s, HIV-2/SIVs and SIVAGMs respectively. Consensus sequences and AACC consensus-like patterns are given with their respective alignments. Protein processing sites are annotated when known. The reader should consult the introduction to Part II for further explanation of the presentation and annotation of the amino acid sequences.

Part III. Analysis. This section is open-ended with the constraint that the sequence analyses and compilations be basic and of interest to the diversity of users. In 1995, the analyses focus upon hybrid (mosaic or recombinant) sequences. Summaries of viral and cellular proteins and of sequencing primers are also included.

Part IV. Related Sequences. Heretofore, this section of the compendium has featured HIV related viral sequences—of nonprimate lentiviruses and the human T-cell lymphotropic viruses. Beginning in 1993, Part IV entries include, with greater emphasis, coding sequences for cellular proteins involved with HIV pathogenesis. Entries are presented in the same format as Part I entries.

Part V. Database Communications. This part contains a printed supplemental reference list for citations in 1994 and 1995. It also provides diskettes of sequences and information about accessing sequences through Internet. The floppy diskettes contain new nucleic acid sequences from Part I (especially those found in alignments) and Part IV and their translated amino acid sequences. For the most current information regarding database files, see the READ.ME file on each diskette. Nucleotide entries are presented in GenBank format for North American users and in EMBL format for European users (unless otherwise requested). Similarly, amino acid sequences are in either PIR or Swiss-Prot format. The diskettes themselves are either 3.5" IBM-DOS format or 3.5" Macintosh format, depending upon what has been requested. If there is any trouble using these files with software designed to work with the format we have sent, please let us know the name of the program you are using and the file that it could not handle.

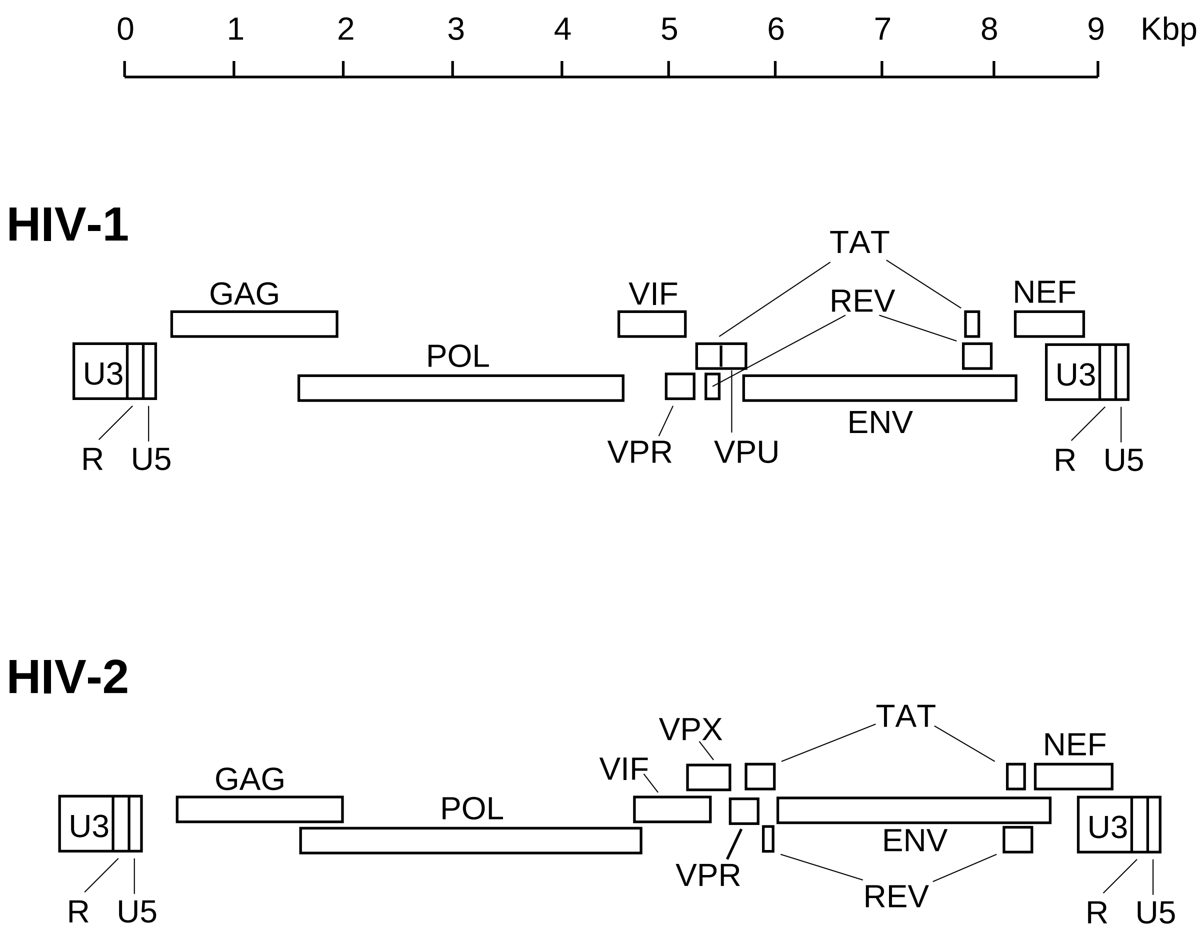
A comprehensive compilation of the nucleic acid and protein sequences published in the Human Retroviruses and AIDS Database since 1987 is available through our Web site, *http://hiv-web.lanl.gov* and on our FTP Server, as described in Part V.

We are prepared to quickly enter both protein and nucleotide sequences into the Human Retroviruses and AIDS database, and in the case of nucleotide sequences, oversee their entry into the large gene libraries. Submission of unpublished sequences is invited and encouraged. Sequence data or inquiries regarding the database should be addressed to

Gerald Myers
Theoretical Division
T-10, MS K710
LANL
Los Alamos, NM 87545

(505)-665-0480; fax (505)-665-3493
e-mail: glm@t10.lanl.gov

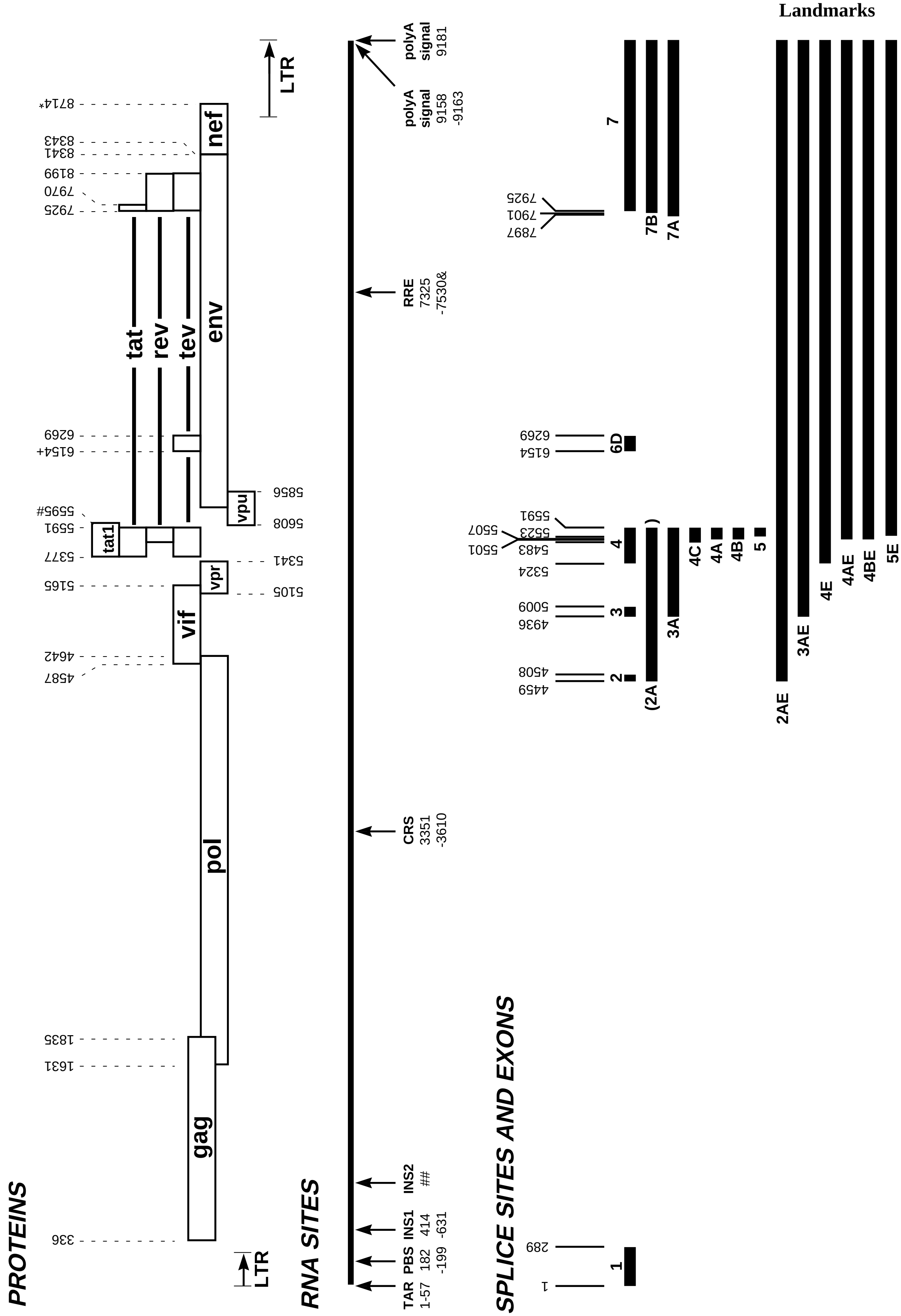
A short glossary follows.


0
1
2
3
4
5
6
7
8
9
Kbp
HIV-1
TAT
GAG
VIF
REV
NEF
U3
POL
U3
ENV
R
U5
VPR
VPU
R
U5
HIV-2
TAT
VPX
NEF
VIF
GAG
POL
U3
U3
ENV
VPR
REV
R
U5
R
U5

HIV/SIV PROTEINS

NAME	SIZE	FUNCTION	LOCALIZATION
Gag MA	p17	membrane anchoring; env interaction; nuclear transport of viral core. (myristylated protein)	virion
CA	p24	core capsid	virion
NC	p7	nucleocapsid, binds RNA	virion
	p6	binds Vpr	virion
Protease (PR)	p15	gag/pol cleavage and maturation	virion
Reverse transcriptase (RT), RNase H	p66 p51 (heterodimer)	reverse transcription, RNase H activity	virion
Integrase (IN)		DNA provirus integration	virion
Env	gp120/gp41	external viral glycoproteins bind to CD4 receptor	plasma membrane, virion envelope
Tat	p16/p14	viral trancriptional transactivator	primarily in nucleolus/nucleus
Rev	p19	RNA transport, stability and utilization factor (phosphoprotein)	primarily in nucleolus/nucleus shuttling between nucleolus and cytoplasm
Vif	p23	promotes virion maturation and infectivity	cytoplasm (cytosol, membranes) virion
Vpr	p10-15	promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M	virion, nucleus (nuclear membrane?)
Vpu	p16	promotes extracellular release of viral particles; degrades CD4 in the ER; (phosphoprotein only in HIV-1 and SIVcpz)	integral membrane protein
Nef	p27-p25	CD4 downregulation (myristylated protein)	plasma membrane, cytoplasm (virion?)
Vpx	p12-16	vpr homolog? (not in HIV-1, only in HIV-2 and SIV)	virion (nucleus?)
Tev	p28	tripartite tat-env-rev protein (also named Tnv)	primarily in nucleolus/nucleus

LANDMARKS ON THE HIV-1 GENOMIC RNA



LANDMARKS:

HIV GENOMIC STRUCTURAL ELEMENTS

LTR Long terminal repeat, the DNA sequence flanking the genome of integrated proviruses. It contains important regulatory regions, especially those for transcription initiation and polyadenylation.

TAR Target sequence for viral transactivation, the binding site for Tat protein and for cellular proteins; consists of approximately the first 45 nucleotides of the viral mRNAs in HIV-1 (or the first 100 nucleotides in HIV-2 and SIV.) TAR RNA forms a hairpin stem-loop structure with a side bulge; the bulge is necessary for Tat binding and function.

RRE Rev responsive element, an RNA element encoded within the env region of HIV-1. It consists of approximately 200 nucleotides (positions 7327 to 7530 from the start of transcription in HIV-1.) The RRE is necessary for Rev function; it contains a high affinity site for Rev; in all, approximately seven binding sites for Rev exist within the RRE RNA. Other lentiviruses (HIV-2, SIV, visna, CAEV) have similar RRE elements in similar locations within env, while HTLVs have an analogous RNA element (RXRE) serving the same purpose within their LTR; RRE is the binding site for Rev protein, while RXRE is the binding site for Rex protein. RRE (and RXRE) form complex secondary structures, necessary for specific protein binding.

CRS cis-acting repressive sequences postulated to inhibit structural protein expression in the absence of Rev. One such site was mapped within the pol region of HIV-1. The exact function has not been defined; splice sites have been postulated to act as CRS sequences.

INS Inhibitory/Instability RNA sequences found within the structural genes of HIV-1 and of other complex retroviruses. Multiple INS elements exist within the genome and can act independently; one of the best characterized elements spans nucleotides 414 to 631 in the gag region of HIV-1. The INS elements have been defined by functional assays as elements that inhibit expression posttranscriptionally. Mutation of the RNA elements was shown to lead to INS inactivation and up regulation of gene expression.

GENES AND GENE PRODUCTS

GAG genomic region encoding the capsid proteins (group specific antigens). The precursor is the p55 myristylated protein, which is processed to p17 (MAtrix), p24 (CApsid), p7 (NucleoCapsid), and p6 proteins, by the viral protease. Gag associates with the plasma membrane where the virus assembly takes place. The 55 kDa Gag precursor is called assemblin to indicate its role in viral assembly.

POL the genomic region encoding the viral enzymes protease, reverse transcriptase and integrase. These enzymes are produced as a Gag-pol precursor polyprotein, which is processed by the viral protease; the Gag-pol precursor is produced by ribosome frameshifting at the C-terminus of gag.

ENV viral glycoproteins produced as a precursor (gp160) and processed to the external glycoprotein gp120 and the transmembrane glycoprotein gp41. The mature proteins are held together by non-covalent interactions; as a result, a substantial amount of gp120 is released in the medium. gp120 contains the binding site for the CD4 receptor.

TAT Transactivator of HIV gene expression. One of the two necessary viral regulatory factors (Tat and Rev) for HIV gene expression. Two forms are known, Tat-1exon (minor form) of 72 amino acids and Tat-2exon (major form) of 86 amino acids. The electrophoretic mobility of these two forms in SDS gels is anomalous, with apparent sizes of approximately 16 kD and 14 kD for Tat- 2exon and Tat-1exon, respectively. Low levels of both proteins are found in persistently infected cells. Tat has been localized primarily in the nucleolus/nucleus by immunofluorescence. It acts by binding to the TAR RNA element and activating transcription initiation and/or elongation from the LTR promoter. It is the first eukaryotic transcription factor known to interact with RNA rather than DNA and may have similarities with prokaryotic anti- termination factors. Extracellular Tat can be found and can be taken up by cells in culture.

REV The second necessary regulatory factor for HIV expression. A 19 kD phosphoprotein, localized primarily in the nucleolus/nucleus, Rev acts by binding to RRE and promoting the nuclear export, stabilization and utilization of the viral mRNAs containing RRE. Rev is considered the most functionally conserved regulatory protein of lentiviruses. Rev cycles rapidly between the nucleus and the cytoplasm.

VIF Viral infectivity factor, typically 23 kD. Promotes the infectivity but not the production of viral particles. In the absence of Vif the produced viral particles are defective, while the cell-to-cell transmission of virus is not affected significantly. Found in almost all lentiviruses, Vif is a cytoplasmic protein, existing in both a soluble cytosolic form and a membrane-associated form. The latter form of Vif is a peripheral membrane protein that is tightly associated with the cytoplasmic side of cellular membranes. Some recent observations suggest that Vif is incorporated in the virion.

VPR Vpr (viral protein R) is a 96-amino acid (14 kd) protein, which is incorporated into the virion. It interacts with the p6gag part of the Pr55gag precursor. Vpr detected in the cell is localized to the nucleus. Proposed functions for Vpr include the nuclear import of preintegration complexes, cell growth arrest, transactivation of cellular genes, and induction of cellular differentiation. Found in HIV-1, HIV-2, SIVmac and SIVmnd. It is homologous to VPX of SIVagm.

VPU Vpu (viral protein U) is unique to HIV-1 and SIVcpz, a close relative of HIV-1. There is no similar gene in HIV-2 or SIV. Vpu is a 16-kd (81-amino acid) type I integral membrane protein with at least two different biological functions: (a) degradation of CD4 in the endoplasmic reticulum, and (b) enhancement of virion release from the plasma membrane of HIV-1-infected cells. Vpu probably possesses an N-terminal hydrophobic membrane anchor and a hydrophilic moiety. It is phosphorylated by casein kinase II at positions Ser52 and Ser56. Vpu is involved in env maturation; not found in the virion.

NEF (previously named 3′ ORF) is an approximately 27-kd myristylated protein produced by an ORF located at the 3′ end of the primate lentiviruses. Other forms of Nef are known, including nonmyristylated variants. Nef is predominantly cytoplasmic and associated with the plasma membrane via the myristyl residue linked to the conserved second amino acid (Gly). Nef has also been identified in the nucleus and found associated with the cytoskeleton in some experiments. Its association with the virion is suspected but not proven. One of the first HIV proteins to be produced in infected cells, it is the most immunogenic of the accessory proteins. Initially thought to be a negative factor, Nef was found to be important for viral replication in vivo. The nef genes of HIV and SIV are dispensable in vitro, but are essential for efficient viral spread and disease progression in vivo. Nef is necessary for the maintenance of high virus loads and for the development of AIDS in macaques. Nef downregulates CD4, the primary viral receptor, and is also proposed to increase viral infectivity. Nef contains PxxP motifs that bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of HIV but not for the downregulation of CD4.

VPX Virion protein of 12 kD found only in HIV-2/SIVagm and not in HIV-1 or SIVmnd. Vpx function in relation to Vpr is not fully elucidated. Vpx is necessary for efficient replication of SIV in PBMCs. Some studies indicate that Vpx and Vpr proteins may be functionally distinct. Progression to AIDS and death in SIV-infected animals can occur in the absence of Vpr or Vpx. Double mutant virus lacking both vpr and vpx was severely attenuated, whereas the single mutants were not, suggesting a redundancy in the function of Vpr and Vpx related to virus pathogenicity.

TEV (also named tnv) tripartite 28 kD viral phosphoprotein produced by some HIV-1 strains. Found primarily in the nucleolus/nucleus. Tev contains the first exon of Tat, a small part of Env and the second exon of Rev. It exhibits both Tat and Rev functions and can functionally replace both essential regulatory proteins of HIV-1. It is produced very early in infection.

STRUCTURAL PROTEINS/VIRAL ENZYMES The products of gag, pol and env genes, which are essential components of the retroviral particle.

REGULATORY PROTEINS Tat and Rev proteins of HIV/SIV and Tax and Rex proteins of HTLVs. They modulate transcriptional and posttranscriptional steps of virus gene expression and are essential for virus propagation.

ACCESSORY OR AUXILIARY PROTEINS additional virion and non-virion- associated proteins produced by HIV/SIV retroviruses: Vif, Vpr, Vpu, Vpx, Nef. Although the accessory proteins are in general not necessary for viral propagation in tissue culture, they have been conserved in the different isolates; this conservation and experimental observations suggest that their role in vivo is very important.

COMPLEX RETROVIRUSES Retroviruses regulating their expression via viral factors and expressing additional proteins (regulatory and accessory) essential for their life cycle.